MESIN PENCARI DOKUMEN DENGAN PENGKLASTERAN SECARA OTOMATIS

Entin Martiana, Nur Rosyid, Usmaida Agusetia
Politeknik Elektronika Negeri Surabaya-Institut Teknologi Sepuluh Nopember
Kampus ITS Keputih Sukolilo Surabaya 60111, Indonesia
Tel:+62-31-5947280 Fax:+62-31-5946114
e-mail: entin@eepis-its.edu, rosyid@eepis-its.edu, usmaida_it04@yahoo.com

Abstract

Web mining in searching based on keywords by automatic clustering is a document searching method by classifying documents based on its keyword. Following is the clustering by centroid linkage hierarchical method (CLHM) to the number of keywords from each document. In clustering, initialization is commonly required for the number of cluster to be formed first, however, in some clustering cases, the user cannot determine how many clusters can be built. Therefore, on this paper, the Valley tracing method is applied as a constraint which identifies variants movement from each cluster formation step and also analyzes its pattern to form automatic clustering. Document data used are from text mining process on documents. Based on 424 documents, this research shows that clustering method using CLHM algorithm can be generally used to classifying documents with exact number automatically.

Keywords*: automatic clustering, CLHM, text mining, valley tracing*

Abstrak

Web mining untuk pencarian berdasarkan kata kunci dengan pengklasteran otomatis adalah suatu metode pencarian dokumen dengan cara mengelompokkan atau mengklaster dokumen dari dokumen-dokumen berdasarkan kata kuncinya. Selanjutnya dilakukan pengklasteran dengan metode centroid linkage hierarchical method (CLHM) terhadap jumlah kata kunci yang diperoleh dari masing-masing dokumen. Dalam pengklasteran, umumnya harus dilakukan inisialisasi jumlah klaster yang ingin dibentuk terlebih dahulu, padahal pada beberapa kasus pengklasteran, user bahkan tidak tahu berapa banyak klaster yang bisa dibangun. Untuk itu, pada makalah ini diaplikasikan metode Valley Tracing sebagai constraint yang akan melakukan identifikasi terhadap pergerakan varian dari tiap tahap pembentukan klaster dan menganalisa polanya untuk membentuk suatu klaster secara otomatis (automatic clustering). Data yang digunakan adalah data hasil dari proses text mining pada dokumen. Dari percobaan yang dilakukan dengan 424 dokumen hasilnya memberikan simpulan bahwa pada umumnya pencarian dokumen menggunakan teknik pengklasteran dengan algoritma CLHM dapat digunakan untuk mengelompokkan dokumen dengan jumlah yang tepat secara otomatis.

Kata kunci: *automatic clustering, CLHM, text mining, valley tracing*

1. PENDAHULUAN

Perkembangan teknologi dewasa ini khususnya internet berkembang sangat pesat. Hal ini diiringi juga dengan semakin berkembangnya Teknologi Informasi yang dibutuhkan oleh pengguna sehingga mengakibatkan munculnya suatu cabang ilmu baru dalam teknologi informasi, yaitu pencarian informasi (*information retrieval*) [1]. Aplikasi pencarian informasi (pencarian dokumen) yang telah ada salah satunya adalah web mining untuk pencarian berdasarkan kata kunci dengan teknik pengklasteran (*clustering*). Pada aplikasi pencarian dokumen sebelumnya, sistem mengelompokkan dokumen dengan menggunakan algoritma K-means, yaitu membangkitkan titik baru secara acak yang nantinya akan digunakan sebagai titik pusat klaster baru sehingga akan terbentuk beberapa klaster sesuai dengan jumlah yang ditentukan. Meskipun sudah menggunakan optimasi K-means, tetapi sistem yang dibangun

belum mampu membentuk jumlah klaster dokumen secara tepat dan otomatis, sehingga dibutuhkan suatu solusi untuk permasalahan tersebut.

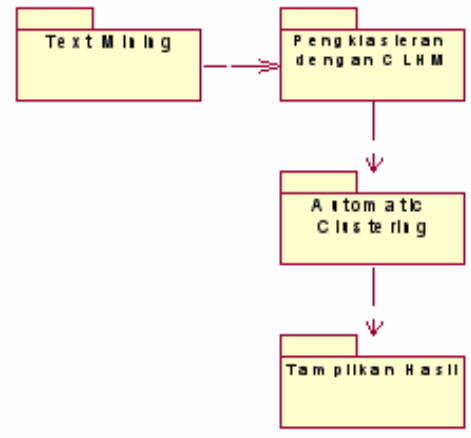
Pada penelitian ini akan dibangun aplikasi *web mining* sebagai mesin pencari berdasarkan kata kunci dengan *automatic clustering* untuk mengelompokkan dokumen secara otomatis. Pada pengklasteran, umumnya harus dilakukan inisialisasi jumlah klaster yang diinginkan terlebih dahulu [2-8], padahal pada beberapa kasus user tidak tahu berapa banyak klaster yang bisa dibangun. Karena itu, pada penelitian ini diaplikasikan metode *Valley tracing* yang mampu menyelesaikan masalah tersebut. Metode ini bisa melakukan pengklasteran secara otomatis (*automatic clustering*) terhadap jumlah kata kunci dari hasil *text mining* yang dilakukan pada dokumen dengan memanfaatkan *centroid linkage hierarchical method* (CLHM), yaitu mendeteksi pergerakan varian pada tiap tahap pembentukan klasternya untuk menemukan *global optimum,* sehingga bisa dibangun klaster secara otomatis [2]. Nilai *centroid* masing-masing hasil pengklasteran selanjutnya akan digunakan untuk menentukan hasil pencarian dokumen.

2. METODE PENELITIAN

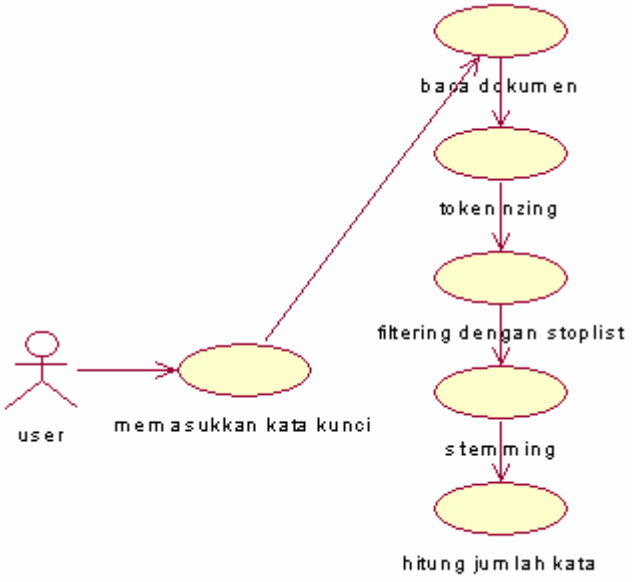
Penelitian ini ditujukan untuk menghasilkan alat bantu untuk mencari dokumen sesuai keinginan dengan memasukkan input berupa kata kunci diantara banyaknya dokumen yang ada. Diagram sistem dari aplikasi pencarian dokumen ini adalah seperti Gambar 1.

2.1. Text Mining

Langkah awal pada penelitian ini adalah pengimplementasian fungsi *text mining* yang meliputi [4]: proses *tokenizing, filtering* dan *stemming,* yang dioptimalkan sebagai sarana untuk pencarian dokumen. Diagram *use case* dari proses *text mining* adalah seperti Gambar 2.



Gambar 1. Garis besar diagram sistem



Gambar 2. Diagram *use case* proses *text mining*

2.2. Pengklasteran Dokumen

Langkah kedua pada penelitian ini adalah proses pengelompokkan dokumen dengan menggunakan CLHM, sesuai langkah berikut [6]:

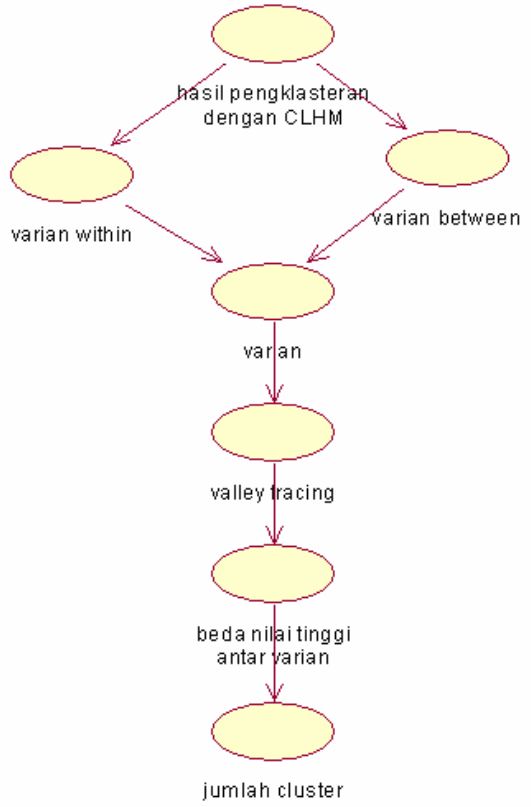
1. Diasumsikan setiap data dianggap sebagai klaster. Kalau n=jumlah data dan c=jumlah klaster, berarti ada c=n.
2. Menghitung jarak antar klaster dengan jarak *Euclidian*.
3. Mencari 2 klaster yang mempunyai jarak *centroid* antar klaster yang paling minimal dan digabungkan ke dalam klaster baru (sehingga c=c-1).

4. Kembali ke langkah 3, dan diulangi sampai dicapai klaster yang diinginkan.
5. Penghitungan jarak antar obyek, maupun antar klasternya dilakukan dengan jarak *Euclidian*, khususnya untuk data numerik [2]. Untuk data 2 dimensi, digunakan persamaan (1).

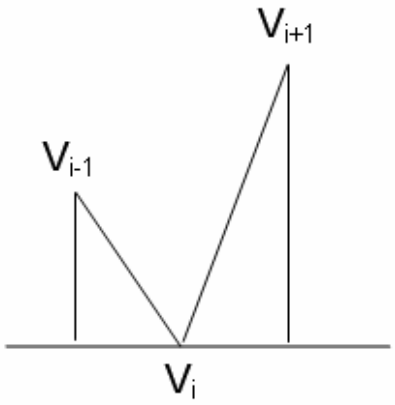
$$d(x,y) = \sqrt{\sum_{i=1}^{n} | x_i - y_i |^2} \quad (1)$$

2.3. Pengklasteran Secara Otomatis

Langkah ketiga adalah menganalisa pola varian untuk mendapatkan posisi *global optimum* dari pola *valley tracing* yang mungkin untuk menentukan jumlah klaster yang tepat secara otomatis sesuai Gambar 3.



Gambar 3. Diagram Alur Proses pengklateran otomatis



Gambar 4. Pola Nilai Beda *Valley-tracing*

2.3.1. Analisa Klaster

Analisa klaster bisa diperoleh dari kepadatan klaster yang dibentuk (*cluster density*). Kepadatan suatu klaster dapat ditentukan dengan *variance within cluster* (Vw) dan *variance between cluster* (Vb). Varian tiap tahap pembentukan klaster dihitung dengan persamaan (2).

$$Vc^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n} (y_i - y_c)^2 \quad (2)$$

dengan: Vc^2 = varian pada klaster c
c = 1..k, dimana k = jumlah klaster
nc = jumlah data pada klaster c
yi = data ke-i pada suatu klaster
yc = rata-rata dari data pada suatu klaster

Selanjutnya dari nilai varian tersebut dhitung nilai *variance within cluster* (Vw) sesuai persamaan (3), sedangkan nilai variance between cluster (Vb) dengan persamaan (4).

$$Vw = \frac{1}{N-c}\sum_{i=1}^{c}(ni-1).Vi^2 \quad (3)$$

$$Vb = \frac{1}{c-1}\sum_{i=1}^{c} n_i(\bar{y}_i - \bar{y})^2 \quad (4)$$

dengan: N = Jumlah semua data
ni = Jumlah data klaster i
Vi = Varian pada klaster i
$\bar{y}$ = rata-rata dari $\bar{y}i$

Salah satu metode yang digunakan untuk menentukan klaster yang ideal adalah batasan varian, yaitu dengan menghitung kepadatan klaster berupa *variance within cluster* (Vw) dan *variance between cluster* (Vb) [7]. Klaster yang ideal mempunyai Vw minimum yang merepresentasikan *internal homogenity* dan maksimum Vb yang menyatakan *external homogenity*.

2.3.2. Valley Tracing

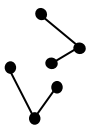
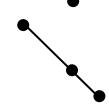
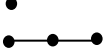
Pada Valley-tracing didefinisikan bahwa kemungkinan mencapai global optimum terletak pada tahap ke-i, jika memenuhi persamaan (5). Persamaan ini diperoleh berdasar analisa pergerakan varian pola *Valley-tracing.*.

$$(v_{i-1} \geq v_i) \cap (v_{i+1} > v_i) \quad (5)$$

dengan: i = 1…n, dan n tahap terakhir pembentukan klaster

Tabel 1 menunjukkan pola-pola *Valley-tracing* yang mungkin mencapai *global optimum*. Pola yang mungkin ditandai dengan simbol √.

Tabel 1. Tabel kemungkinan pola Valley-tracing mencapai global optimum [8]

Pola	Mungkin	Pola	Mungkin
	√		X
	√		X
	√		X
	X		√
	X		X
	X		X
	X		X
	X		X

Selanjutnya, dengan pendekatan metode *valley-tracing* dilakukan identifikasi perbedaan nilai tinggi (*∂*) pada tiap tahap dengan persamaan (6). Nilai *∂* digunakan untuk menghindari local optima, dimana persamaan ini diperoleh dari maksimum *∂* yang dipenuhi pada persamaan (6). Untuk membentuk klaster secara otomatis, yaitu klaster yang mencapai global optima, digunakan nilai λ sebagai threshold, sehingga klaster secara otomatis terbentuk ketika memenuhi persamaan (7).

$$\partial = (V_{i+1} - V_i) + (V_{i-1} - V_i) = (V_{i+1} + V_{i-1}) - (2 \times V_i) \quad (6)$$

$$\max(\partial) \geq \lambda \quad (7)$$

Guna mengetahui keakuratan dari suatu metode pembentukan klaster pada hierarchical method, dengan menggunakan valley-tracing digunakan persamaan (8), dengan nilai terdekat ke max(*∂*) adalah nilai kandidat max(*∂*) sebelumnya. Nilai φ yang lebih besar atau sama dengan 2 (φ≥2), menunjukkan klaster yang terbentuk merupakan klaster yang well-separated (terpisah dengan baik).

$$\varphi = \frac{\max(\partial)}{nilai\ terdekat\ ke\ \max(\partial)} \quad (8)$$

Tahap terakhir adalah proses *sorting* untuk pengurutan dokumen yang memiliki kata kunci paling banyak sampai yang paling sedikit.

3. HASIL DAN PEMBAHASAN

Aplikasi *web mining* untuk pencarian berdasarkan kata kunci dengan pengklasteran otomatis ini diterapkan untuk pencarian dokumen berdasarkan inputan kata kunci dari dokumen yang bertema “lumpur lapindo” dengan jumlah 60 dokumen dengan ektensi *.txt yang hasilnya dibandingkan dengan metode pencarian dengan menggunakan algoritma K-means yang telah dioptimasi.

3.1. Uji Ketepatan dari Jumlah Klaster Dokumen yang Terbentuk

Uji coba ini digunakan untuk mengetahui ketepatan dari jumlah klaster dokumen yang terbentuk secara otomatis dalam pencarian dokumen dengan menggunakan metode CLHM. Kata kunci yang dimasukkan: “tanggul jebol”. Pada Gambar 5 ditunjukkan hasil keluaran dari mesin pencari yang dikembangkan. Dari pengujian ini didapatkan bahwa jumlah klaster yang terbentuk secara otomatis adalah empat klaster. Jumlah klaster yang dianggap optimal merupakan tahap pembentukan klaster yang mempunyai nilai beda tinggi yang terbesar atau maksimal. Sedangkan Gambar 6 menunjukkan grafik pergerakan pola varian dari tahap pembentukan klaster, sedangkan grafik nilai beda tingginya adalah seperti Gambar 7. Karena jumlah klaster dianggap optimal jika tahap pembentukan klaster mempunyai nilai beda tinggi yang terbesar atau maksimal, maka jumlah klaster yang terbentuk secara otomatis adalah tepat, yaitu pada tahap pembentukan jumlah klaster sebanyak 4.

3.2. Uji Pembandingan Hasil Pencarian Dokumen

Uji coba ini digunakan untuk membandingkan hasil pencarian dokumen antara menggunakan metode K-means dan CLHM dengan jumlah klaster yang sama, yaitu empat klaster. Percobaan ini dilakukan dengan memasukkan kata kunci yang hampir sama, yaitu kata kunci: ”tindakan warga” pada sistem pencarian dengan menggunakan metode K-means dan CLHM. Hasil penghitungan jumlah kata kunci pada masing-masing dokumen dan hasil pencarian dokumennya ditunjukkan pada Tabel 2.

Dari pengujian ini diketahui bahwa dengan inputan kata kunci yang sama, yaitu ”tindakan warga”, proses pengklasteran dokumen dengan menggunakan metode K-means dan CLHM memperoleh hasil pencarian dokumen yang yang hampir sama juga, yaitu dokumen yang ada sebagai hasil pencarian dengan metode CLHM juga merupakan hasil pencarian dengan metode K-means. Hal ini disebabkan karena jumlah kata kunci dari masing-masing

dokumen adalah sama, hanya saja dalam CLHM data dikelompokkan menjadi jumlah *klaster* secara otomatis, sedangkan dalam metode K-means pengelompokkan data bergantung pada titik pusat awal *klaster* yang telah dioptimasi, sehingga mempengaruhi kedekatan jarak antar data.

:: Web Mining untuk Pencarian Berdasarkan Kata Kunci Dengan Automatic Clustering::

String:

Reset Generate

Tokenizing
tanggul
jebol

Stop List
tanggul
jebol

Stemming
tanggul
jebol

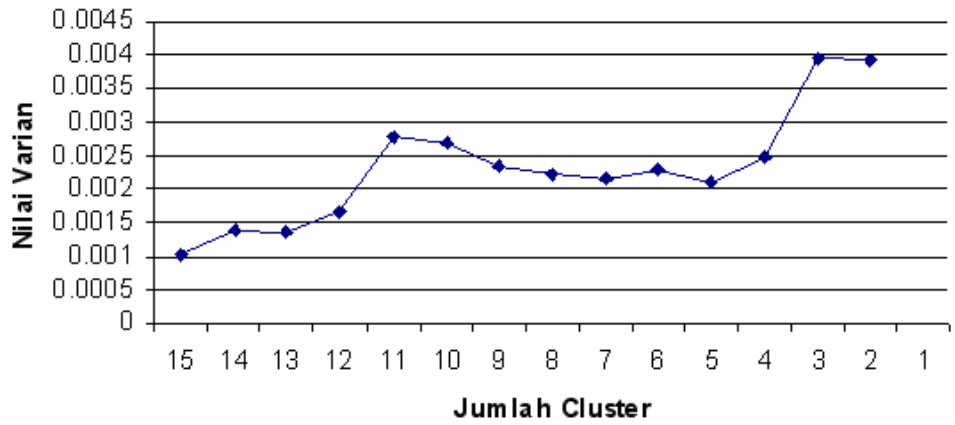
Keyword counter & Automatic Cluster
Number Of Cluster: 4

No.	File Name	Kata Kunci ke-1	Kata Kunci ke-2	Cluster
1	01 - Tempo Interaktif-01 (Korban Minta Presiden Ambil Alih Bencana Lapindo)_stem_hitung.txt	0	0	**4**
2	02 - Tempo Interaktif-02 (Interpelasi Lapindo Akan Diajukan Hari Ini)_stem_hitung.txt	0	0	**4**
	……			
	.....			
423	Tanah Ambles Terdeteksi Tiga Pekan Lalu_stem_hitung.txt	2	0	**4**
424	Tanggul Porak Poranda Semburan Lumpur Mengganas_stem_hitung.txt	9	3	**3**

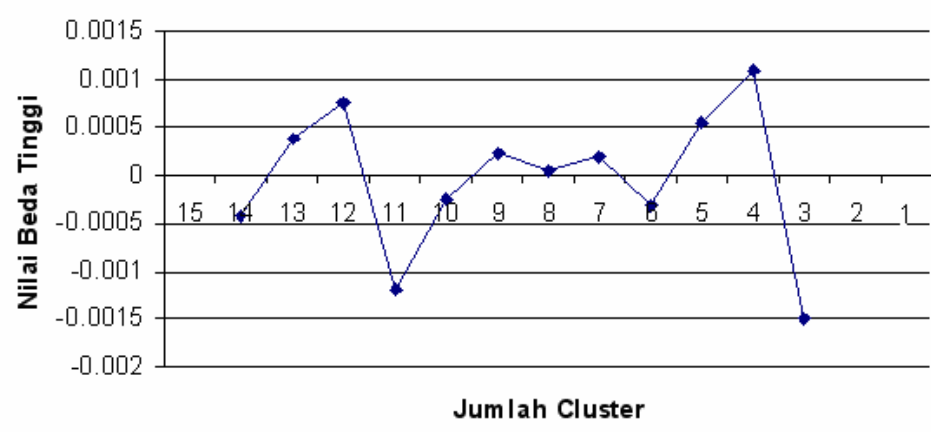
Besarnya Data Dalam Setiap Cluster
Cluster ke-1: 141.125
Cluster ke-2: 18.693877551020407
Cluster ke-3: 82.0625
Cluster ke-4: 0.25277008310249305

Hasil Automatic Cluster
Number Of Cluster: 4, Member: 1
1. 51 - Pusat Lumpur Lapindo Meledak.txt
2. 28 - Wikipedia-01 (Banjir lumpur panas Sidoarjo).txt
3. 53 - Rovicky-01 (Banjir lumpur panas Sidoarjo).txt
4. 40 - Lumpur Panas Berbalik Arah.txt

Gambar 5. Hasil pengujian pencarian dokumen



Gambar 6. Grafik Pergerakan Pola Varian



Gambar 7. Grafik Nilai Beda Tinggi

Tabel 2. Tabel perbandingan hasil pencarian dokumen dengan metode K-means dan CLHM

Hasil Pencarian Dokumen dengan K-means	Hasil Pencarian Dokumen dengan CLHM
1. 47 - Pengungsi Lumpur Panas Terserang ISPA - 26-06-2006, 1125 WIB - KOMPAS Cyber Media - NASIONAL.txt	1. 47 - Pengungsi Lumpur Panas Terserang ISPA - 26-06-2006, 1125 WIB - KOMPAS Cyber Media - NASIONAL.txt
2. 08 - Tempo Interaktif-08 (Korban Lapindo Blokir Jalan).txt	2. 08 - Tempo Interaktif-08 (Korban Lapindo Blokir Jalan).txt
3. 17 - Tempo Interaktif-17 (7 Korban Lapindo Derita Gangguan Jiwa).txt	3. 17 - Tempo Interaktif-17 (7 Korban Lapindo Derita Gangguan Jiwa).txt
4. 35 - Hot Mud Flow-01 (Luas bangunan korban lusi).txt	4. 35 - Hot Mud Flow-01 (Luas bangunan korban lusi).txt
5. 33 - Dua Warga Korban Lumpur Panas Sidoarjo Meninggal.txt	
6. 54 - Semburan Baru di Rumah Penduduk.txt	

3.3. Uji Pembandingan Waktu Kinerja

Pengujian ini digunakan untuk membandingkan waktu kinerja yang dibutuhkan untuk pencarian dokumen pada sistem pencarian dokumen dengan menggunakan pengklasteran metode K-means dan CLHM. Kata kunci yang digunakan: “warga kebanjiran”. Waktu eksekusi dari metode K-means dan CLHM berturut-turut adalah 5 menit 30 detik dan 6 menit 12 detik. Dari pengujian ini dapat diketahui bahwa proses pengklasteran dokumen dengan

menggunakan metode CLHM memerlukan waktu yang lebih lama jika dibandingkan dengan pengklasteran dengan menggunakan metode K-means. Hal ini disebabkan karena dalam CLHM data tidak langsung dikelompokkan kedalam beberapa klaster dalam satu tahap, tetapi dimulai dari satu klaster yang mempunyai jarak yang dekat, dan berjalan seterusnya selama beberapa iterasi, hingga terbentuk beberapa klaster tertentu.

4. SIMPULAN

Pencarian dokumen menggunakan teknik pengklasteran dengan algoritma CLHM dan analisa pola varian yang memenuhi *valley tracing* dapat digunakan untuk mengelompokkan dokumen dengan jumlah *klaster* yang tepat secara otomatis, meskipun memerlukan waktu komputasi yang lebih lama.

DAFTAR PUSTAKA

[1]. Agus AZ, Setiono AN. *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*. Proceeding of SITIA. Surabaya. 2002: 1-6.
[2]. Barakbah AR, Arai K. *Determining Constraints of Moving Variance to Find Global Optimum and Make Automatic Clustering*. Proceeding of IES. Surabaya. 2004: 409-413.
[3]. Uramoto N, Matsuzawa H, Nagano T, Murakami A, Takeuchi H, Takeda K. A Text-Mining System for Knowledge Discovery from Biomedical Documents. *IBM Systems Journal*. 2004; 43(3): 516-533.
[4]. Hammouda KM, Kamel MS. Efficient phrase-based document indexing for Web document clustering. *Knowledge and Data Engineering, IEEE Transactions on*. 2004; 16(10): 1279-1296.
[5]. Bulacu M, Schomaker L. Text-Independent Writer Identification and Verification Using Textural and Allographic Features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2007; 29(4): 701-717.
[6]. Ashraf F, Ozyer T, Alhajj R. Employing Clustering Techniques for Automatic Information Extraction From HTML Documents. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*. 2008; 38(5): 660-673.
[7]. Man L, Chew Lim T, Jian S, Yue L. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2009; 31(4): 721-735.
[8]. Barakbah AR, Arai K. *Identifying Moving Variance to Make Automatic Clustering for Normal Data Set*. In. Proc. IECI Japan Workshop (IJW). Tokyo. 2004: 125-134.